

Gap-Measure Tests with Applications to Data Integrity Verification

Truc Le^{*1}, Jeffrey Uhlmann²

Department of Computer Science, University of Missouri - Columbia, 201 EBW, Columbia, MO, USA

^{*}tdlxqb@mail.missouri.edu; ²uhlmannj@missouri.edu

Abstract

In this paper we propose and examine gap statistics for assessing uniform distribution hypotheses. We provide examples relevant to data integrity testing for which max-gap statistics provide greater sensitivity than chi-square (χ^2), thus allowing the new test to be used in place of or as a complement to χ^2 testing for purposes of distinguishing a larger class of deviations from uniformity. We establish that the proposed max-gap test has the same sequential and parallel computational complexity as χ^2 and thus is applicable for Big Data analytics and integrity verification.

Keywords

Hypothesis Testing; Distribution Testing; Chi-Square Testing; Data Integrity; Big Data; Gap Statistics; Max Gap; Min Gap; Data Integrity; Gonzalez Algorithm; Closest Pair

Introduction

Distribution testing is a fundamental statistical problem that arises in a wide range of practical applications. At its core, the problem is to assess whether a dataset that is assumed to comprise samples from a known probability distribution is in fact consistent with that assumption. For example, if the end state of a computer simulation of a physical system is a set of points with an expected physics-prescribed distribution, then any detected deviation from that expected distribution could undermine confidence in the results obtained and possibly in the integrity of the simulation system itself.

Data integrity verification is a related application for distribution testing in which the objective is to detect evidence of tampering, e.g., human-altered data. For example, many sources of numerical data produce numbers with first digits conforming to the Benford-Newcomb first-digit distribution (This phenomenon is often referred to as “Benford’s Law”) [1, 2], while digits other than the first and last are uniformly sampled from $\{0, \dots, 9\}$ [3]. Digits in human-created numbers, by contrast, tend to exhibit high regularity with all elements of $\{0, \dots, 9\}$ represented with nearly equal cardinality. Statistically identified deviations of this kind have been used to uncover acts of scientific misconduct and accounting fraud [4, 5, 6, 7, 8, 9], but there is an increasing need for higher-sensitivity tests.

There is of course no way to make an unequivocal binary assessment of whether a dataset of samples conforms to a given distribution assumption, but it is possible to devise statistical tests which can assign a rigorous likelihood estimate to the hypothesis that the dataset does (or does not) represent samples from the assumed distribution. In this paper we briefly review the most widely-used method for distribution testing, the chi-square (χ^2) test, and then develop alternative tests based on the statistics of gap-widths between data items of consecutive rank. Our principal contribution is a max-gap test which is shown to provide superior sensitivity to regularity deviations from a uniform distribution that are relevant to data integrity testing [10, 11, 12]. We show that this test can be evaluated with the same optimal computational complexity (serial and parallel) as the conventional χ^2 test and is therefore suitable for extremely large-scale datasets.

Chi-square Test

The χ^2 test is a statistical measure that can be applied to a discrete dataset to assess the hypothesis that its elements were sampled from a particular distribution. More specifically, it is a histogram-based method to measure the goodness-of-fit between the observed frequency distribution and the expected (theoretical) frequency

distribution. The general procedure of the test includes the following steps:

1. Calculate the chi-square statistic, χ^2 , which is a normalized sum of squared differences (deviations) between observed and expected frequencies.
2. Determine the degrees of freedom, df , of that statistic, which is essentially the number of frequencies reduced by the number of parameters of the fitted distribution.
3. Compare χ^2 with the critical value for the chi-square distribution with df degrees of freedom.

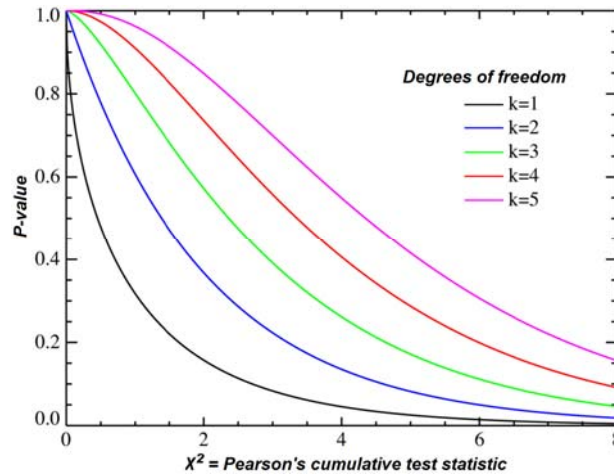


FIG. 1 COMPLEMENT OF THE CUMULATIVE DISTRIBUTION FUNCTION OF THE χ^2 DISTRIBUTION, SHOWING χ^2 ON THE x-AXIS AND P-VALUE ON THE y-AXIS [13]

An example of the complement of the cumulative distribution function of the χ^2 distribution is shown in Fig. 1 with different degrees-of-freedom values. For uniformity testing, the procedure can be expressed as follows:

1. Given N observations, construct an N -bin histogram. Let b_i be the bin count for the i^{th} bin ($i=1, \dots, N$), which is the observed as frequency distribution. As we are testing for uniformity, the expected frequency distribution $e_i = 1, \forall i=1, \dots, N$.
2. Compute the chi-square test statistic:

$$\chi^2 = \sum_{i=1}^N \frac{(b_i - e_i)^2}{e_i} = \sum_{i=1}^N (b_i - 1)^2 \quad (1)$$

3. The number of degrees of freedom, df , is $N - 1$ for this case because the counts for $N - 1$ bins uniquely determine the count for the remaining bin.
4. Compute the complement of the cumulative distribution function of the χ^2 distribution with χ^2 and df obtained from the previous steps. Compare this value with the significance level α for the test result.

Despite being the de-facto standard for assessing dataset consistency with respect to a given distribution assumption, the χ^2 test is not optimally sensitive to the types of deviation from uniformity that arise in many data integrity applications. One example involves narrow-band missing data resulting from a corrupted sensor or measurement process. Another example involves data that are generated from a non-random process and exhibits a higher degree of data regularity than is expected for a uniform distribution [14, 15]. Datasets of the latter kind are typical of artificial and human-generated data, e.g., as in a forged dataset that has been tailored to include deviations that qualitatively resemble (to humans) uniform random deviates. In the following section we demonstrate the advantage of the proposed max-gap test over χ^2 for narrow-band and high-regularity deviations from uniformity.

Max-gap Test

The maximum gap, or max-gap, for a dataset of real values is defined as the maximum difference between

elements of consecutive rank, which can be determined from a sorted ordering of the dataset. The distribution of spacings between consecutive-rank items in a dataset has been examined in the literature [16, 17, 18, 19], and we summarize here some of the results relevant to gap analysis. Assume we are given $N - 1$ observations on the open unit interval $(0, 1)$ which divide the interval into N intervals whose lengths in ascending order are denoted by $S_{(1)} < S_{(2)} < \dots < S_{(N)}$. For uniformity testing, we are interested in $S_{(N)}$, as it is the *max-gap* of the observations. The exact distribution of $S_{(N)}$ is [19]:

$$P(S_{(N)} \leq x) = \sum_{v=0}^N (-1)^v \binom{N}{v} (1 - Nx)_+^{N-1}, \quad (2)$$

where $a_+ = \max(a, 0)$.

From the p-value of the max-gap $S_{(N)}$, denoted by p , we can perform a max-gap test for uniformity by checking the condition $p \geq \alpha$ for the *one-sided* test, or $1 - \frac{\alpha}{2} \geq p \geq \frac{\alpha}{2}$ for the *two-sided* test, where α is the significance level. When N is large, we may replace computation of the exact cumulative distribution of the max-gap in Eqn. 2 with the following asymptotic result [19]:

$$P(S_{(N)} \leq x) \stackrel{N \rightarrow \infty}{=} e^{-e^{\ln N - Nx}}, \quad (3)$$

where the expected value of $S_{(N)}$ is

$$E(S_{(N)}) \stackrel{N \rightarrow \infty}{=} \frac{\gamma + \ln N}{N}, \quad (4)$$

where γ is Euler's constant.

An efficient max-gap test for uniformity can then be formalized as follows: Given $N - 1$ observations x_i , and a significance level α , compute the max-gap $S_{(N)}$ of $\{0, 1\} \cup \{x_i\}$. Next, the p-value of the statistics is calculated as:

$$p = 1 - e^{-e^{\ln N - NS_{(N)}}} \quad (5)$$

If the p-value satisfies $p \geq \alpha$ for the one-sided test, or $1 - \frac{\alpha}{2} \geq p \geq \frac{\alpha}{2}$ for the two-sided test, the observations are deemed to pass the test. Otherwise, the set of observations is assessed to be inconsistent with a uniform-sampling hypothesis and fails the test.

In the next section, we present results of experiments comparing the relative sensitivities of the χ^2 test and the max-gap test for, e.g., indentifying anomalous regularity in a presumed-uniform distribution.

Experiments

In this section, we compare the max-gap test versus the most well-known and commonly used χ^2 test. We conducted four experiments involving datasets of $N = 10,000$ samples, with the result for each experiment obtained as an average of one million independent tests. Sensitivity is assessed by comparing the respective p-values for the one-sided forms of the two tests, where smaller values indicate greater sensitivity. The first experiment was performed using a dataset of samples from a true uniform distribution. As expected, the dataset passed both tests for uniformity with $p = 0.5$.

The second experiment examined sensitivity to the difference between a uniform distribution and a normal distribution with standard deviation σ sampled within a fixed interval $(0, 1)$. The distinctive shape of the normal distribution is realized within the interval when σ is small but flattens with increasing values and approaches uniformity. Both tests are equally sensitive for small σ , and both approach $p = 0.5$ for large σ , but the χ^2 test exhibits higher sensitivity for intermediate values (see Fig. 2). The latter is not surprising because the χ^2 test is ideally sensitive to deviations from normality.

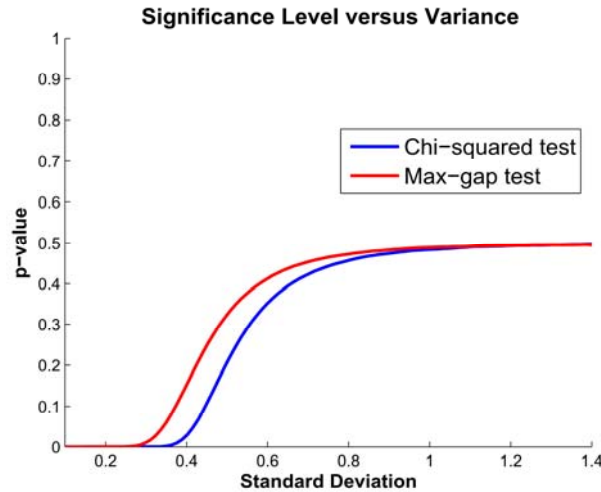


FIG. 2 THE P-VALUES OF THE χ^2 TEST AND THE MAX-GAP TEST OF A NORMAL DISTRIBUTION SAMPLED WITHIN A FIXED INTERVAL. WHEN THE STANDARD DEVIATION (σ) IS SMALL, BOTH TESTS EASILY IDENTIFY THE DATA'S NON-UNIFORMITY. AS σ INCREASES, THE DATA DISTRIBUTION APPROACHES UNIFORMITY WITHIN THE SAMPLE INTERVAL AND HENCE THE P-VALUES CONVERGE TO 0.5. THIS IS AN EXAMPLE IN WHICH THE χ^2 TEST PROVIDES INHERENTLY GREATER SENSITIVITY THAN THE MAX-GAP TEST

The third experiment examined sensitivity of the two tests to a uniform distribution with a narrow-band exclusion (Fig. 3). This of course is a problem for which the max-gap test is ideally suited, and Eqn. 4 reveals that superior sensitivity. What is possibly the most interesting about the results is that the χ^2 test provides only modest sensitivity even as the exclusion width approaches one percent of the distribution window.

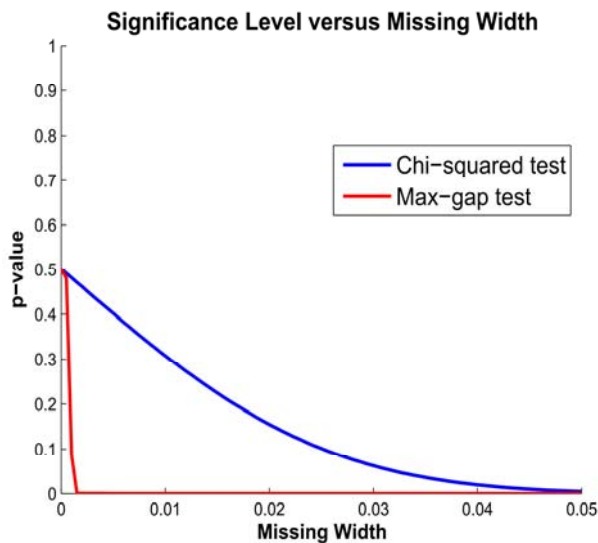


FIG. 3 P-VALUES OF THE χ^2 TEST AND THE MAX-GAP TEST FOR NARROW-BAND MISSING DATA. IN THIS CASE THE MAX-GAP TEST PROVIDES INHERENTLY GREATER SENSITIVITY

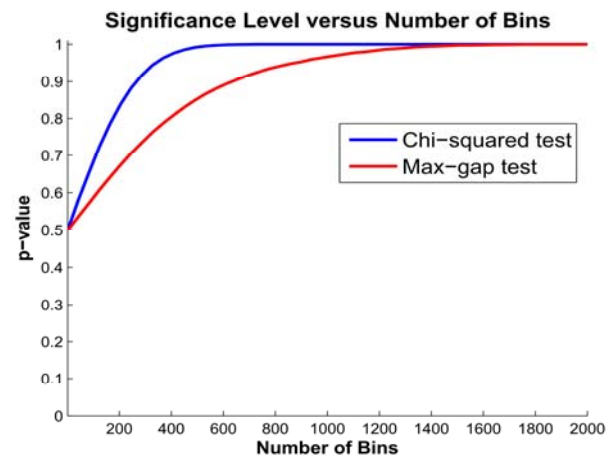


FIG. 4 P-VALUES OF THE χ^2 TEST AND THE MAX-GAP TEST FOR HIGH REGULARITY DATA. REGULARITY FOR A DATASET OF SIZE N IS PARAMETERIZED BY A NUMBER OF BINS K WITH N/K UNIFORM SAMPLES WITHIN EACH OF K -EQUAL BINS. THUS $K = 1$ GENERATES A UNIFORM DISTRIBUTION AND INCREASING K APPROACHES REGULAR SPACING. THE MAX-GAP TEST AGAIN DEMONSTRATES GREATER SENSITIVITY

The fourth experiment is the most relevant to data integrity applications. It examined sensitivity to regularity in sample spacing. Anomalous distribution regularity is a common characteristic of human-altered data because people typically underestimate the degree of natural "clustering" that is present in data sampled from a truly uniform distribution. As a consequence, human-created or human-altered data tend to have higher regularity, i.e., tend to be "more evenly distributed", than what is expected for uniformly-distributed data. More generally, high-regularity deviations from uniformity can arise from the unanticipated influence of a structured or non-random process, e.g., frequency-combing effects from a physical sensor or simulation artifacts resulting from a low-quality pseudorandom number generator.

A regularity parameter $1 \leq k \leq N$ was used for this experiment by uniformly distributing N/k samples within each of k equal-width subdivisions of the distribution interval. Thus, $k = 1$ represents a uniform sampling over the entire interval and produces a uniform distribution; and as k increases to N , the spacing between samples becomes increasingly regular. Although uniform and high-regularity distributions are difficult for humans to distinguish visually, Fig. 4 shows that the max-gap test provides significantly higher sensitivity than χ^2 to subtle regularity deviations from uniformity.

Min-Gap Test

The one-sided variants of the max-gap and χ^2 tests were used because they provide a practical balance between high sensitivity and low false alarm rates, but the one-sided or two-sided of either test may provide the optimal trade-off for the needs of a particular given application. In some applications, the optimal trade-off might be obtained from a *min-gap*, $S_{(1)}$, test. The min-gap approximated distribution is given by [19]

$$P(S_{(1)} \leq x) \stackrel{N \rightarrow \infty}{=} e^{-e^{\ln N - Nx}} \sum_{v=0}^{N-1} \frac{(e^{\ln N - Nx})^v}{v!}, \quad (6)$$

and its expected value is [19]

$$E(S_{(1)}) \stackrel{N \rightarrow \infty}{=} \frac{\gamma + \ln N + \sum_{i=1}^{N-1} \frac{1}{i}}{N} \quad (7)$$

A min-gap test can be defined and performed analogously to the max-gap test and would be ideally suited for detecting spuriously-replicated data items. However, several simpler non-statistical methods can be applied to detect replicated data, so the potential applications of the min-gap test may be somewhat more limited than the max-gap test.

Computational Considerations

In terms of computational complexity, both χ^2 and max-gap tests can be evaluated in optimal $O(N)$ time and $O(N)$ space. This complexity is achieved for max-gap by the use of the Gonzalez algorithm [20, 21] to determine the max-gap in linear time without sorting. The Gonzalez algorithm performs a special binning which guarantees by the pigeonhole principle that the max-gap data items will be found as the maximum and minimum values, respectively, in consecutive non-empty bins. This algorithm allows the max-gap test to be evaluated in optimal $O(N)$ time and space, i.e., the same as χ^2 , and is as efficiently parallelizable as the χ^2 test (The max-gap and χ^2 tests are both highly amenable to parallelization with $O(N/P)$ time complexity on P processors.).

The min-gap pair needed to implement a min-gap test which can be identified in optimal expected $O(N)$ time and space using Rabin's randomized closest-pair algorithm [22, 23]. Unlike the Gonzalez algorithm for max-gap, Rabin's algorithm generalizes efficiently to higher dimensions.

Discussion and Future Work

We have defined and developed a max-gap test for distinguishing deviations from uniformity in a 1D dataset of size N . By using Gonzalez's algorithm, we have shown that this test can be performed with commensurate efficiency, both serial and in parallel, with the conventional χ^2 test. Our experiments demonstrate that the max-gap test provides improved sensitivity in two particular applications of relevance to data integrity verification. More generally, the proposed max-gap and min-gap tests are of potential value as alternatives or to complement the use of χ^2 for distribution testing and discrimination.

There are many statistical tests for equality of distributions beyond the χ^2 test such as the Kolmogorov-Smirnov test [24, 25, 26] and the Cramer-von Mises test [27]. Of course there can be no test that is uniformly superior to all others

for all possible distributions, but it appears that most of the standard tests examined in the literature would be challenged similarly to the χ^2 test to distinguish uniform from regularly distributed data.

Potential future work could consider tests which jointly combine gap and χ^2 statistics into a more sophisticated single test [28] which allows greater flexibility to optimize the sensitivity and false alarm trade-off for problems of high practical interest, e.g., big data analytics and integrity verification. On the algorithmic side, we have pointed out that the Gonzalez algorithm does not generalize to higher dimensions; however, relatively efficient subquadratic algorithms do exist for solving the largest empty circle and largest empty rectangle problems in two dimensions such as the algorithms in [29, 30]. Tests on 2D distributions could also potentially exploit information about the largest empty region of a Voronoi decomposition or the distribution of nearest-neighbor distances from a Delaunay triangulation. In $d > 2$ dimensions, it may be possible to devise gap-related statistical tests based on results from efficient algorithms for identifying approximations to the largest empty d -sphere or d -rectangle, but this is purely speculative. In higher dimensions, it may be better to abandon gap-type statistics and focus on statistics gleaned from efficiently-computable k -d and orthant (quad, octant, etc.) tree decompositions of point sets.

If computational efficiency is less of a concern, a perhaps more fruitful direction for highly-sensitive distribution testing in high dimensions is to examine the length of the Euclidean minimum spanning tree (EMST) for a dataset. The expected length of the EMST of uniformly-distributed points can be determined using analysis similar to what has been described in this paper for estimating the expected values for the max and min gaps in 1D, and we conjecture that EMST length is likely to be more sensitive to many practically important types of deviations from uniformity than the conventional χ^2 test. Such an EMST test would be computationally expensive (though subquadratic), but this cost could be justified in applications for which subtle deviations are critically important, e.g., high-fidelity physics simulations.

REFERENCES

- [1] M. Nigrini and J. Wells, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, ser. Wiley Corporate F&A. Wiley, 2012. [Online]. Available: <http://books.google.com/books?id=FdRPh787I7oC>
- [2] C. Winter, M. Schneider, and Y. Yannikos, "Model-based digit analysis for fraud detection overcomes limitations of benford analysis," *Seventh International Conference on Availability, Reliability and Security*, vol. 0, pp. 255–261, 2012.
- [3] S. Dlugosz and U. Mller-Funk, "The value of the last digit: Statistical fraud detection with digit analysis," *Advances in Data Analysis and Classification*, vol. 3, no. 3, pp. 281–290, 2009. [Online]. Available: <http://dx.doi.org/10.1007/s11634-009-0048-5>
- [4] R. Pirracchio, M. Resche-Rigon, S. Chevret, and D. Journois, "Do simple screening statistical tools help to detect reporting bias?" *Annals of Intensive Care*, vol. 3, no. 1, 2013. [Online]. Available: <http://dx.doi.org/10.1186/2110-5820-3-29>
- [5] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, August 2002. [Online]. Available: <http://dx.doi.org/10.1214/ss/1042727940>
- [6] N. Kingston and A. Clark, *Test Fraud: Statistical Detection and Methodology*, ser. Routledge Research in Education. Taylor & Francis, 2014. [Online]. Available: <http://books.google.com/books?id=3fzpAwAAQBAJ>
- [7] M. Nigrini, *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*, ser. Wiley Corporate F&A. Wiley, 2011. [Online]. Available: <http://books.google.com/books?id=ct9CB4eJCXYC>
- [8] A. Diekmann, *Methodological Artefacts, Data Manipulation and Fraud in Economics and Social Science*, ser. Jahrbucher Nationalokonomie und Statistik. Lucius & Lucius, 2011. [Online]. Available: <http://books.google.com/books?id=vzJlcjAz4sC>
- [9] L. Leemann and D. Bochler, "A systematic approach to study electoral fraud," *Electoral Studies*, vol. 35, no. 0, pp. 33–47, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0261379414000390>
- [10] Y. Fujii, "The analysis of 168 randomised controlled trials to test data integrity," *Anaesthesia*, vol. 67, no. 6, pp. 669–670, 2012. [Online]. Available: <http://dx.doi.org/10.1111/j.1365-2044.2012.07189.x>

- [11] S. Al-Marzouki, S. Evans, T. Marshall, and I. Roberts, "Are these data real? statistical methods for the detection of data fabrication in clinical trials," *BMJ*, vol. 331, no. 7511, pp. 267–270, 2005.
- [12] U. Simonsohn, "Just post it the lesson from two cases of fabricated data detected by statistics alone," *Psychological science*, vol. 24, no. 10, pp. 1875–1888, 2013.
- [13] M. Haggstrom. (2010) Complement of chi-square cumulative distribution. [Online]. Available: http://en.wikipedia.org/wiki/File:Chi-square_distributionCDF-English.png
- [14] J. H. Pitt and H. Z. Hill, "Statistical detection of potentially fabricated data: A case study," *ArXiv e-prints*, November 2013.
- [15] H. Z. Hill and J. H. Pitt, "Failure to replicate: A sign of scientific misconduct?" *Publications*, vol. 2, no. 3, pp. 71–82, 2014. [Online]. Available: <http://www.mdpi.com/2304-6775/2/3/71>
- [16] D. A. Darling, "On a class of problems related to the random division of an interval," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 239–253, June 1953.
- [17] R. Pyke, "Spacings," *Journal of the Royal Statistical Society*, pp. 395–449, 1965.
- [18] R. Pyke, "Spacings revisited," pp. 417–427, 1972.
- [19] L. Holst, "On the lengths of the pieces of a stick broken at random," *Journal of Applied Probability*, pp. 623–634, September 1980.
- [20] T. Gonzalez, "Algorithms on sets and related problems," *Department of Computer Science, University of Oklahoma, Norman, OK, Tech. Rep.*, 1975.
- [21] T. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985.
- [22] M. Golin, R. Raman, C. Schwarz, and M. Smid, "Simple randomized algorithms for closest pair problems," *Nordic Journal of Computing*, vol. 2, no. 1, pp. 3–27, March 1995.
- [23] R. Lipton, "Rabin flips a coin," in *The P=NP Question and Gdels Lost Letter*. Springer US, 2010, pp. 77–80.
- [24] Z. W. Birnbaum and F. H. Tingey, "One-sided confidence contours for probability distribution functions," *The Annals of Mathematical Statistics*, vol. 22, no. 4, pp. 592–596, December 1951. [Online]. Available: <http://dx.doi.org/10.1214/aoms/1177729550>
- [25] W. Conover, *Practical nonparametric statistics*, ser. Wiley series in probability and statistics: Applied probability and statistics. Wiley, 1999. [Online]. Available: <https://books.google.com/books?id=dYEpAQAAAMAAJ>
- [26] G. Marsaglia, W. W. Tsang, and J. Wang, "Evaluating Kolmogorov's distribution," *Journal of Statistical Software*, vol. 8, no. 18, pp. 1–4, 2003. [Online]. Available: <http://www.jstatsoft.org/index.php/jss/article/view/v008i18>
- [27] H. Cramer, "On the composition of elementary errors," *Scandinavian Actuarial Journal*, vol. 1928, no. 1, pp. 13–74, 1928. [Online]. Available: <http://dx.doi.org/10.1080/03461238.1928.10416862>
- [28] D. Maynes, "Combining statistical evidence for increased power in detecting cheating," in *Presentation given at the annual meeting of the National Council of Measurement in Education, San, Diego, CA, April, 2009*.
- [29] B. Chazelle, R. L. Drysdale, and D. T. Lee, "Computing the largest empty rectangle," *SIAM Journal of Computing*, vol. 15, no. 1, pp. 300–315, February 1986.
- [30] A. Naamad, D. Lee, and W. Hsu, "On the maximum empty rectangle problem," *Discrete Applied Mathematics*, vol. 8, no. 3, pp. 267–277, 1984.